L'AED et SAS/INSIGHT, Visualisations dynamiques des données

Monique Le Guen

L'AED et SAS/INSIGHT
Visualisations dynamiques des données

Monique Le Guen
CNRS - MATISSE – UMR 8595[1]
leguen@univ-paris1.fr

Résumé
Par opposition aux tests d'hypothèses destinés à vérifier des hypothèses a priori sur des relations entre variables, l'Analyse Exploratoire des Données est utilisée pour rechercher et découvrir des relations systématiques entre variables, en prenant en compte un grand nombre de variables.
Dans les techniques de l'exploratoire, la visualisation et les techniques informatiques de manipulations dynamiques des données via les graphiques jouent un rôle majeur.
Depuis une décennie des logiciels orientés AED s'appuyant sur les capacités de la micro-informatique et l'interactivité entre les fenêtres donnent au statisticien praticien des outils et de nouvelles stratégies d'analyse des données.
Ces techniques visuelles changent le « paysage » du statisticien praticien comme celui de l'apprenant (novice en statistiques).
Dans notre présentation-démonstration interactive nous essayerons de montrer l'apport du module SAS/Insight, pour un utilisateur SAS novice comme pour un utilisateur SAS expérimenté en statistiques.

Mots Clés
Analyse Exploratoire des données, Graphiques Exploratoires, SAS/INSIGHT.

Abstract *Exploratory Data Analysis and SAS/Insight: Dynamic Visual Display of the Data*

Contrary to Tests of Hypotheses intended to verify *a priori* hypotheses regarding relations between variables, Exploratory Data Analysis (EDA) is used to search and discover systematic relationships between variables by taking into count a large number of variables. In these exploratory techniques, visual displays and dynamic manipulations of data using graphics are essential. For the past decade, along with the advent of new capabilities in computer sciences and the interactivity between windows, the EDA-oriented software provides the skilled statistician with new tools and strategies to analyse data.
These visual techniques change the "landscape" for professional statisticians as well as novice learners in statistics.
In our interactive demonstration and presentation, you will learn more about SAS/Insight and how this module can help a novice SAS user, as well as an experienced SAS user in statistics.

Key Words
Exploratory Data Analysis, Exploratory Graphics, SAS/INSIGHT.

Codes JEL : C49 ; C87.

[1] MATISSE-CNRS UMR8595, Maison des Sciences Economiques, 106-112 Boulevard de l'Hôpital, 75013 Paris.

Sommaire

Introduction

Par opposition aux tests d'hypothèses traditionnels destinés à vérifier des hypothèses a priori sur des relations entre variables, l'Analyse Exploratoire des Données est utilisée pour rechercher et découvrir des relations systématiques entre variables, en prenant en compte un grand nombre de variables.

Les techniques de l'Exploratoire regroupent à la fois :

- les méthodes élémentaires d'analyse de distribution et
- les méthodes d'Analyse Exploratoire Multidimensionnelle des données.

Dans les techniques de l'exploratoire, la visualisation et les techniques informatiques de manipulations dynamiques des données via les graphiques jouent un rôle majeur.

Depuis une décennie des logiciels orientés AED s'appuyant sur les capacités de la micro-informatique et l'interactivité entre les fenêtres donnent au statisticien praticien des outils et de nouvelles stratégies d'analyse des données .
Ces techniques visuelles changent le « paysage » du statisticien praticien comme celui de l'apprenant (novice en statistiques).

Dans notre présentation-démonstration interactive nous essayerons de montrer l'apport du module SAS/Insight, pour un utilisateur SAS novice en statistiques comme pour un utilisateur SAS expérimenté.

L'Analyse Exploratoire des Données

L'Analyse Exploratoire AED en français ou EDA en anglais est apparue dans les années 1970 à l'Université de Princeton (USA) sous la plume et l'enseignement de TUKEY (1977). Dans les pays anglophones l'AED s'est alors répandue dans la communauté des praticiens de la Statistique comme dans celle des enseignants. Les enseignants en SHS et en techniques d'Ingénieurs ont été les premiers à reconnaître son efficacité pédagogique (ERIKSON & NOSANCHUK 1979, VELLEMAN 1980, MARSH C. 1988). En France l'AED est peu connue et peu enseignée.

TUKEY (1980) a proposé de voir l'Analyse de données au sens large de Data Analysis comme une coopération entre l'Analyse Exploratoire des données et l'Analyse Confirmatoire (« We need Both Exploratory and Confirmatory Data Analysis »).

Dans le protocole d'analyse des données tel que le définit TUKEY, exploration puis confirmation, l'approche confirmatoire suit une **logique d'expérimentation**. Elle met en oeuvre les techniques de la statistique mathématique classique, avec ses briques de base que sont la moyenne, la loi normale, le test de Student, l'égalité des variances, la linéarité, la modélisation. Ces techniques sont rigoureuses, bien formalisées, impressionnantes même. Mais reposant sur des présupposés, elles ne sont plus optimales si ces derniers ne sont pas vérifiés.

L'approche exploratoire suit au contraire une **logique d'observation**. L'explorateur va regarder ses données sous différentes facettes, tenter de mettre en évidence les structures qu'elles recèlent, enfin et le cas échéant formuler des hypothèses plausibles.

Les méthodes exploratoires élémentaires sont proches de celles de la statistique descriptive – recherche d'observations atypiques, asymétrie d'une distribution, non normalité, bimodalité etc. - mais complétées par de nouveaux indicateurs plus résistants et par des techniques plus robustes.

Les méthodes exploratoires multivariées portent sur les relations entre de nombreuses variables et s'attachent à aider l'œil à révéler et à « *mieux voir ce que les données ont à nous dire* ».

Les méthodes d'analyse des données à la Française (BENZECRI & CO) ainsi que les techniques neuronales comme les cartes de Kohonen viennent logiquement s'insérer dans cette démarche de l'Analyse Exploratoire Multidimensionnelle (LEBART, MORINEAU & PIRON, et COTTRELL M. & LETREMY P.).

Les Pièges de la Modélisation

Montrons les pièges de la régression linéaire, sur un exemple tiré de l'ouvrage de TOMASSONE, LESQUOY, MILLIER (1986) "*La Régression nouveaux regards sur une ancienne méthode statistique*" Masson.
Cet exemple construit pour le propos est inspiré de l'exemple original de F.J . ANSCOMBE "*Graphs in Statistical Analysis*". C'était en 1973 avant que l'analyse des résidus devienne un standard d'études incontournable pour la validation d'une régression.

Effectuons une régression linéaire sur les 5 couples de variables ci-dessous (X,Ya), (X,Yb), (X,Yc), (X,Yd) et (Xe,Ye).

Variables ➔ 1 2 3 4 5 6 7

OBS	X	Ya	Yb	Yc	Yd	Xe	Ye
1	7	5,535	0,113	7,399	3,864	13,715	5,654
2	8	9,942	3,770	8,546	4,942	13,715	7,072
3	9	4,249	7,426	8,468	7,504	13,715	8,491
4	10	8,656	8,792	9,616	8,581	13,715	9,909
5	12	10,737	12,688	10,685	12,221	13,715	9,909
6	13	15,144	12,889	10,607	8,842	13,715	9,909
7	14	13,939	14,253	10,529	9,919	13,715	11,327
8	14	9,450	16,545	11,754	15,860	13,715	11,327
9	15	7,124	15,620	11,676	13,967	13,715	12,746
10	17	13,693	17,206	12,745	19,092	13,715	12,746
11	18	18,100	16,281	13,893	17,198	13,715	12,746
12	19	11,285	17,647	12,590	12,334	13,715	14,164
13	19	21,365	14,211	15,040	19,761	13,715	15,582
14	20	15,692	15,577	13,737	16,382	13,715	15,582
15	21	18,977	14,652	14,884	18,945	13,715	17,001
16	23	17,690	13,947	29,431	12,187	33,281	27,435

Surprise, les résultats obtenus dans les 5 analyses sont strictement identiques:

- même droite de régression,
- même R^2,
- mêmes erreurs-type sur les coefficients,
- mêmes STUDENT et
- mêmes p-values.

Analyse Statistique

Model Equation				
YA	=	0.52	+	0.81 X

Summary of Fit			
Mean of Response	12.60	R-Square	0.62
Root MSE	3.22	Adj R-Sq	0.59

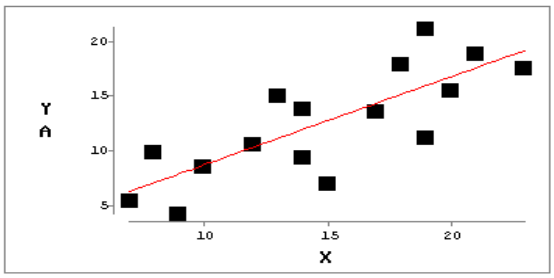
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
Model	1.0	234.6	234.6	22.6	0.0003
Error	14.0	145.4	10.4		
C Total	15.0	380.1			

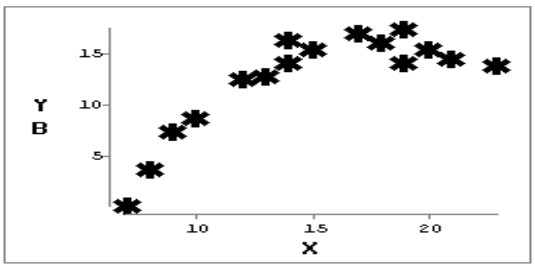
Parameter Estimates							
Variabl	DF	Estimat	Std Err	T Stat	Prob >|	Toleran	Var Inflat
INTERC	1.00	0.52	2.67	0.20	0.8476	.	0
X	1.00	0.81	0.17	4.75	0.0003	1.00	1.00

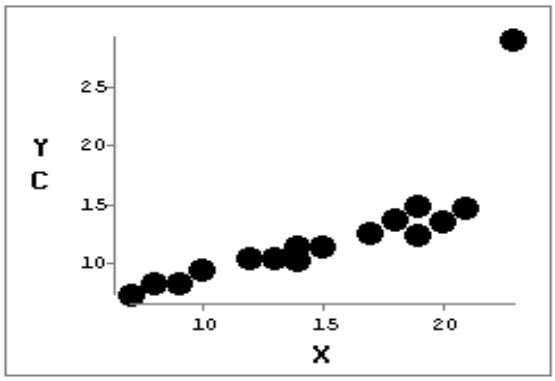
Si l'utilisateur s'arrête à l'analyse des indicateurs standard (R^2, F, T de Student) l'adéquation du modèle aux données peut être jugée digne d'intérêt, aucune contre-indication n'est décelable.

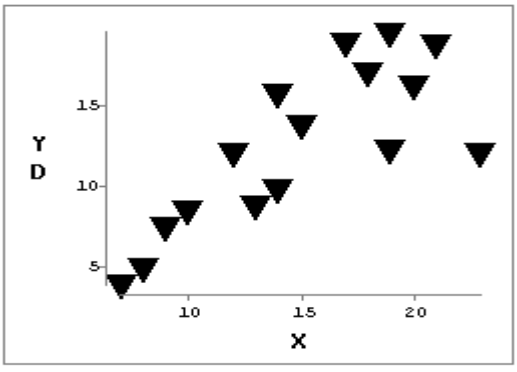
De Simples Nuages de Points

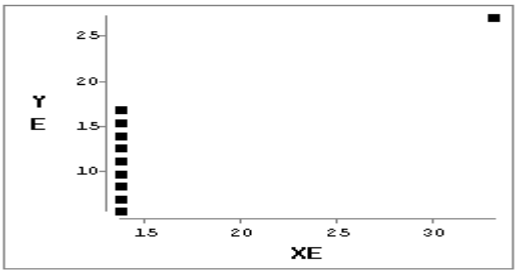
De simples nuages de points, graphiques (X,Y) permettent de vérifier que les suppositions de la régression linéaire ne sont pas vérifiées.











SAS/INSIGHT

SAS/Insight est un module atypique de SAS qui permet grâce à la visualisation et l'interactivité, entre l'utilisateur et la machine d'une part, et l'interactivité entre les fenêtres d'affichage d'autre part, de pouvoir pratiquer et mettre en œuvre certaines techniques de l'Analyse Exploratoire des données.

SAS/Insight offre des possibilités étendues en matière de représentation graphique (histogramme, Bar Chart, Box Plot, Mosaic Plot, Scatter Plot, Line Plot, Rotating Plot, Qqplot, Leverage Plot, Biplot etc.) Tous ces graphiques sont interactifs et peuvent être manipulés et enrichis grâce à une boîte à outils comprenant, un curseur, une main, un zoom, des couleurs, des marqueurs, pour repérer des points observations répondant à un certain critère de sélection, et des styles de ligne pour les graphiques de courbes et de séries chronologiques.

Un tableur, pour créer ou afficher des tables SAS, permet de saisir, corriger, déplacer, extraire des données et ré-exprimer des variables (plus de trente fonctions mathématiques et statistiques sont proposées).

SAS/Insight calcule des statistiques résistantes, par exemple des moyennes tronquées ou winsorisées qui viennent s'ajouter aux statistiques descriptives classiques paramétriques ou non paramétriques. SAS/Insight inclut des fonctions de mise en oeuvre de procédures d'ajustement paramétriques (ajustement de la distribution de la variable étudiée selon la loi normale, la loi log-normale, la loi exponentielle ou encore la loi de Weibull), mais également non paramétriques (estimations de probabilité selon la méthode du noyau).

Dans l'esprit multivarié anglo-saxon, quatre méthodes existent : *principal component rotation analysis (Y's), canonical correlation analysis (X's,Y's), maximum redundancy analysis (X's,Y's), canonical discriminant analysis (X's,Y nominal),* sont venues compléter depuis la version 8, l'unique méthode « multivariate », Analyse en Composantes Principales (ACP). Ont également étaient ajoutés des graphiques Biplot (2D et 3D) pour représenter simultanément des variables et des observations sur un même graphique.

Cependant le praticien français reste très démuni dans sa pratique habituelle des données multidimensionnelles. On ne trouve dans SAS/Insight aucune procédure liée à l'Analyse des Correspondances (simple ou multiple) ou à la classification.
L'utilisateur français pourra cependant récupérer les sorties d'une ACP, AFC ou ACM réalisées par les procédures SAS et représenter les plans factoriels avec SAS/Insight en les enrichissant par des couleurs et des symboles pour faciliter l'analyse des plans factoriels.
Exemple : Repérer, Colorier sur les plans factoriels les contributions des points à un axe factoriel, supérieures à une certaine valeur.

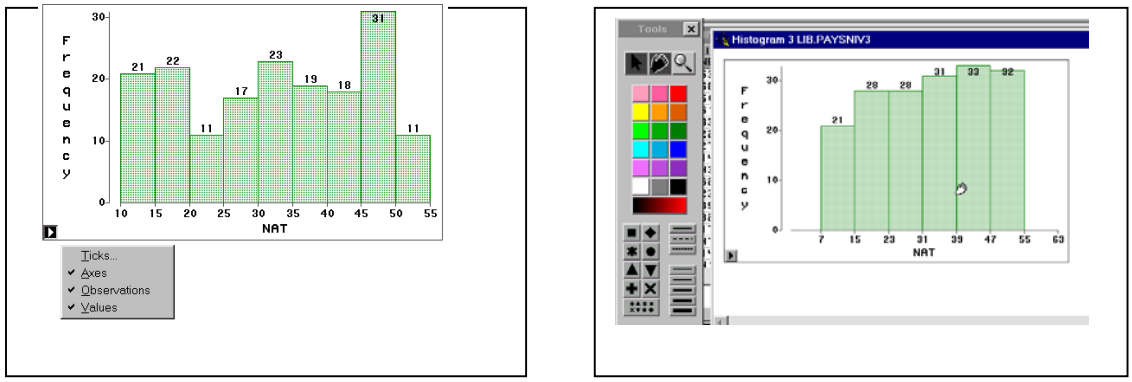
SAS/Insight permet également de pratiquer de manière conviviale et même ludique l'analyse confirmatoire. Il intègre un menu de modélisation **Fit** très complet, permettant de réaliser un grand nombre de modèles comme la régression linéaire

avec les indicateurs de BKW (Belsey, Kuh et Welsh) pour diagnostiquer les observations influentes, les problèmes de multi-colinéarités, le modèle linéaire généralisé, la régression logistique, l'analyse de Variance, Covariance, les modèles Probit et Logit.

Des Exemples de Graphiques

Nous présentons ici quelques graphiques standards. L'objet de notre présentation-démonstration sera de montrer une gamme plus importante de graphiques et de manipulations.
Ces graphiques sont obtenus via des menus et avec quelques clics de souris. Aucune ligne de code n'est à écrire !

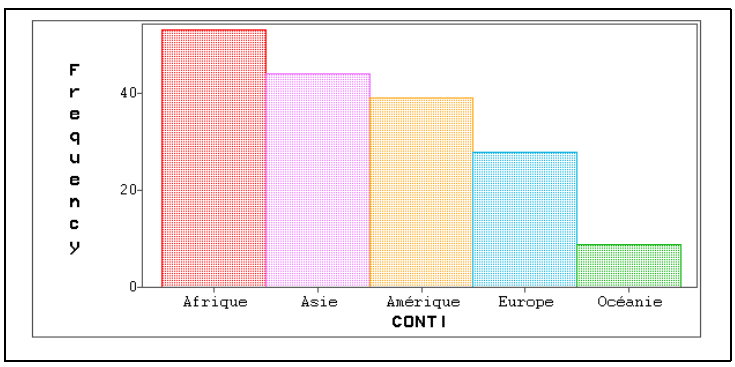
Histogramme



L'affichage de deux histogrammes de la même variable Nat (Taux de Natalité) avant et après changement des intervalles de classes par la souris, met sous les yeux de l'utilisateur l'effet de sa manipulation. Il peut ainsi découvrir que : *l'histogramme n'est pas un estimateur robuste de la densité de probabilité*.

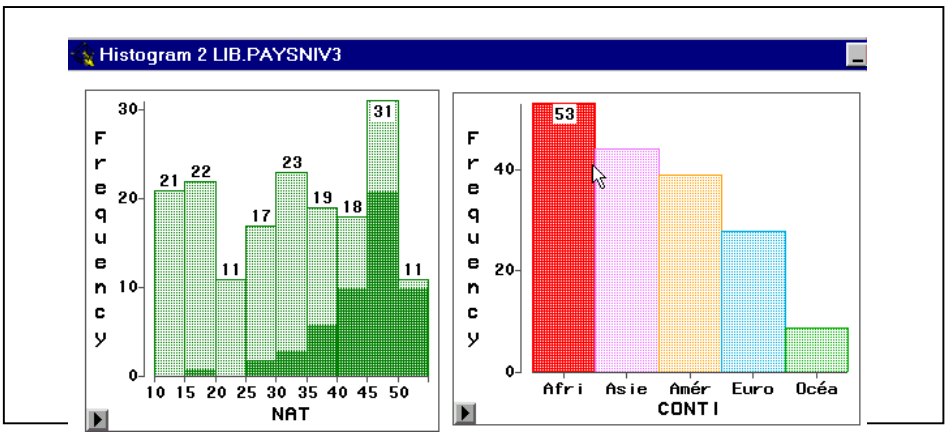
Bar Chart

Le bar Chart est un histogramme en couleur pour une variable nominale (qualitative).



Interactivité entre les graphiques

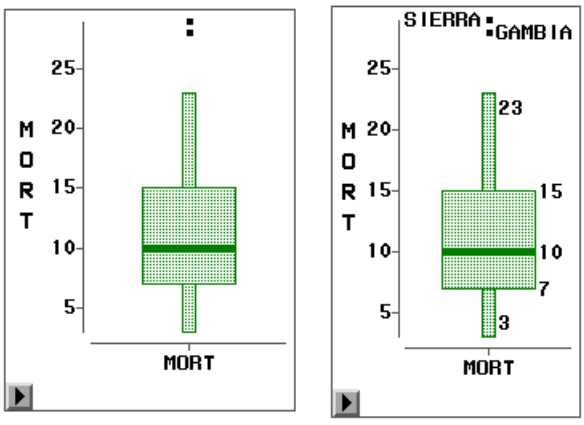
L'affichage simultané de l'histogramme de la variable NAT et le Bar Chart de la variable CONTI apporte une nouvelle dimension.



En sélectionnant sur le Bar Chart les observations de l'Afrique, la distribution des pays africains vient s'afficher dans la distribution globale, tous continents confondus. En cliquant sur les pays européens la distribution viendrait s'afficher du côté des taux de natalité faible. L'utilisateur voit immédiatement que pour ces deux continents les distribution des taux de natalité sont à l'opposé.

Le Box Plot de Tukey

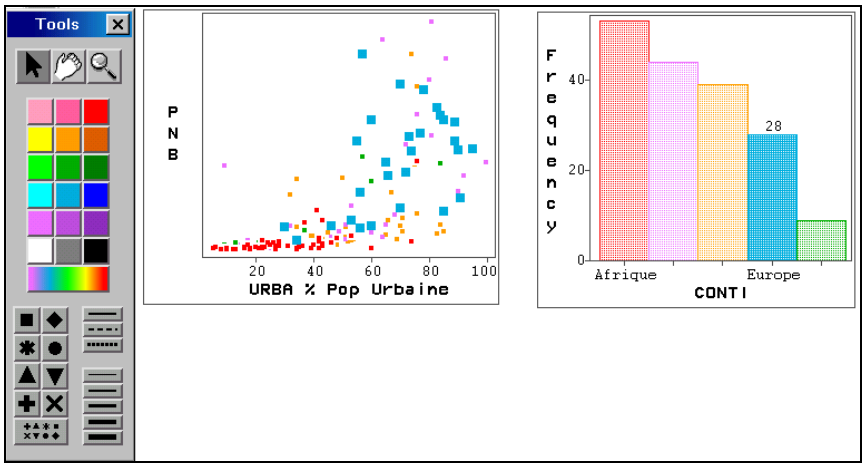
Le Box plot de TUKEY est une représentation de la distribution d'une variable à partir



des quantiles. Il permet de repérer les points atypiques, de les identifier par une étiquette, d'apprécier visuellement l'asymétrie de la distribution et plus encore (Le Guen, 2002) .

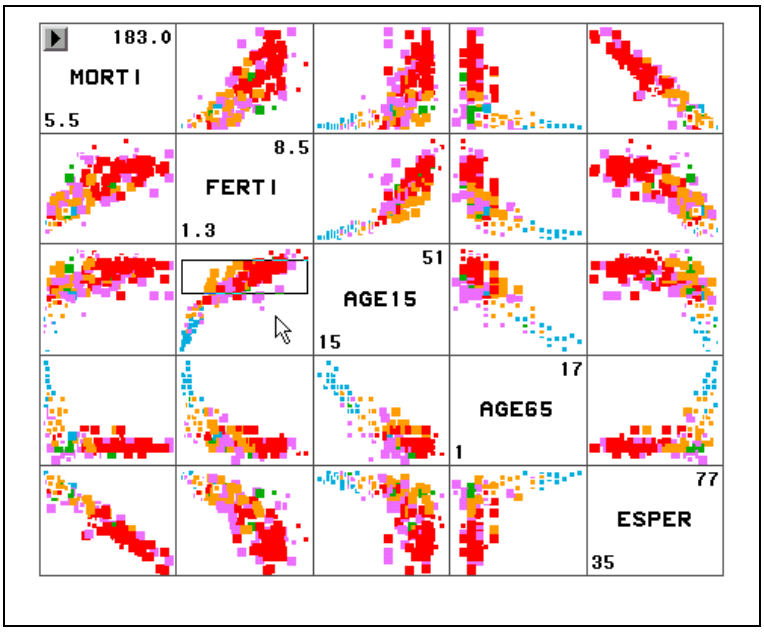
Diagrammes de Dispersion

Par l'intermédiaire de la boîte à outils, le graphique cartésien du PNB en fonction du taux d'urbanisation (URBA), vient s'enrichir par affectation des couleurs du Bar Chart aux points observations des pays.



Brushing et Slicing dans une matrice de Diagrammes de Dispersion

La matrice de diagrammes de dispersion est une représentation visuelle de la matrice de corrélation. Les liaisons linéaires entre les variables sont rapidement détectées, ainsi que les autres formes de liaisons.



Le *brushing* et le *slicing* sur un diagramme de dispersion sont des techniques exploratoires de brossage des données qui reposent sur les capacités d'interactivité du logiciel, interactivité homme-machine et interactivité (Link) entre les fenêtres d'affichage (BECKER, CLEVELAND, (1987).
L'utilisateur sélectionne par l'intermédiaire d'un rectangle « brosse », dessiné avec la souris, un sous-ensemble des observations affichées.
Toutes les observations sélectionnées par la « brosse » apparaissent en surbrillance dans les autres graphiques. L'utilisateur peut changer, la taille de la « brosse », la déplacer, faire des tranches de valeurs (slicing), pour une exploration dynamique des relations entre plus de 2 variables.
SAS/Insight ne peut pas se décrire avec des mots, il faut le voir à l'œuvre.

Conclusion

La visualisation et l'interactivité des logiciels apportent un « paysage » nouveau pour le praticien de l'analyse des données, comme pour l'apprenant d'une technique statistique. C'est l'orientation des recherches actuelles tant au niveau du Datamining (*pour extraire des données ce qu'elles ont à nous dire*) que du côté de l'enseignement de la Statistique. Pour améliorer et rendre plus efficace l'enseignement complexe de la Statistique, les neuro-sciences et la neuro-pédagogie doivent être parties prenantes de ces enjeux.

Une simple métaphore pour conclure:
L'Analyse Exploratoire des Données est au cerveau droit ce que l'Analyse Confirmatoire est au cerveau gauche, les deux doivent communiquer pour traiter l'information. AED et AC ne sont pas opposées, elles sont complémentaires.

Références

- ANSCOMBE F. J. (1973), « Graphs in Statistical Analysis », *American Statistician*, 27.

- BANOS A. (2001), « À propos de l'analyse spatiale exploratoire des données, About Exploratory Data Analysis», *CYBERGEO*, n°197.

- BECKER A., CLEVELAND W. (1987), « Brushing scatterplots », *Technometrics*, 29(2), p. 127-142.

- BELSEY D.A., KUH E., WELSH R.E. (1980), *Regression diagnostics*. New York, Wiley.

- CLEVELAND W. S. (1993), *Visualizing Data*, Hobart Press, Summit, New Jersey, USA.

- CLEVELAND W. S. (1994), *The Elements of Graphing Data*, Hobart Press, Summit, NJ.

- COTTRELL M. (2001), « Analyse de Kohonen, classification et analyse exploratoire de données » in *Actes du VIII^{e} Congrès de la Société Francophone de Classification*, Université Antilles-Guyane, Pointe-à-Pitre, p. 6-7.

- COTTRELL M. (2003), *Les réseaux de neurones : historique, méthodes et applications*, Transparents, 142 p. ftp://samos.univ-paris1.fr/pub/SAMOS/preprints/samos174.pdf

- COTTRELL M. & LETRÉMY P. (2003), *Algorithme de Kohonen : classification et analyse exploratoire des données*. Transparents nouvelle version. 156 pages. ftp://samos.univ-paris1.fr/pub/SAMOS/preprints/samos173.pdf

- DESTANDAU S., LADIRAY D., LE GUEN M. (1999), « Analyse exploratoire des données », *Courrier des statistiques*, juin, n° 90.

- DESTANDAU S., LADIRAY D., LE GUEN M. (1999), « AED mode d'emploi », *Courrier des statistiques*, juin, n^{o} 90, p. 17-21. http://matisse.univ-paris1.fr/leguen/leguen1999a.pdf

- DESTANDAU S. & LE GUEN M. (1998), « Analyse exploratoire des données avec SAS/Insight », INSEE Guides n°7-8.

- ERICKSON, B. H., & NOSANCHUK, T. A. (1992), *Understanding data: an introduction to exploratory and confirmatory data analysis for students in the social sciences*. 2nd Ed. Milton Keynes, Open University Press.

- FOX J. & LONG J.S. (1990), *Modern Methods of Data Analysis*. Sage Pub.

- HOAGLIN, D. C., MOSTELLER, F., & TUKEY, J. W. (1985), *Understanding Robust and Exploratory Data Analysis.* New York, J. Wiley.

- LADIRAY D. (1998), « L'Analyse exploratoire des données (Exploratory Data Analysis) », *Lettre du SSE*, INSEE, n°30, septembre.

- LADIRAY D. (2000), « Graphiquez vos données », *Journal de la Société française de Statistique*, vol. 141, p. 61-67.

- LEBART L., MORINEAU A. & PIRON M. (1995), *Statistique exploratoire multidimentionnelle*. Paris, Dunod.

- LE GUEN M. (1999), « Références en statistiques, sciences cognitives et enseignement », *Courrier des statistiques*, juin, n° 90, p. 39-44. http://matisse.univ-paris1.fr/leguen/leguen1999g.pdf

- LE GUEN M. (2001), « Repenser l'Initiation à la Statistique », *Statistiquement vôtre*, n° 4, site de la SFDS ou http://matisse.univ-paris1.fr/leguen/leguen2001a.pdf

- LE GUEN M. (2002), « La boîte à moustaches de Tukey, un outil pour initier à la Statistique », *Bulletin de méthodologie sociologique*, n° 73, p. 43-64. http://matisse.univ-paris1.fr/leguen/leguen2001b.pdf

- LE GUEN M. (2003), « Tableaux croisés et diagrammes en mosaïque, pour visualiser les probabilités marginales et conditionnelles », *Bulletin de méthodologie sociologique*, n° 77, p. 62-79. http://matisse.univ-paris1.fr/doc2/leguen1491.pdf

- LETREMY P. (2002), « Programmes et macros SAS inspirés de Kohonen (présentés à la journée CLUB SAS de décembre 2002) ». 46 transparents. http://samos.univ-paris1.fr/samos176.pdf

- MARSH C. (1988), *Exploring data: an introduction to data analysis for social scientists*, Cambridge, Polity Press.

- TOMASSONE R., LESQUOY E., MILLIER C. (1986), *La Régression nouveaux regards sur une ancienne méthode statistique.* Paris, Masson.

- TUKEY, J. W. (1977), *Exploratory Data Analysis*. Reading, MA, Addison-Wesley.

- TUKEY, J. W. (1980), *«* We need both exploratory and confirmatory statistics », *The American Statistician*, vol. 34, n° 1, p. 23-25.

- VELLEMAN, P. F. & HOAGLIN, D. C. (1980), *Applications, Basics and Computing of Exploratory Data Analysis.* Boston, Mass., Duxbury Press.

Sites Internet orientés EDA

http://exploringdata.cqu.edu.au/ : ce site fournit des ressources matérielles pour enseigner la statistique dans l'esprit Analyse exploratoire de données.

http://www.itl.nist.gov/div898/handbook/eda/eda.htm
NIST/SEMATECH e-Handbook of Statistical Methods ou Engineering Statistics Handbook

http://www.itl.nist.gov/div898/handbook/graphgal.htm : présente une galerie de graphiques exploratoires issus du *Handbook* précédent. Avec pour chaque type de graphique une définition, des exemples, un descriptif, des questions sur le graphique, des réalisations logicielles, etc.

LADIRAY D., *Macros SAS pour l'Analyse Exploratoire des Données*, http://www.unige.ch/ses/sococ/eda/sas/welcome.html

L'association MIRAGE (Mouvement International pour le développement de la recherche en Analyse graphique et exploratoire) accueille toutes les personnes (enseignants, chercheurs) qui s'intéressent à l'analyse graphique et exploratoire des données. http://www.unige.ch/ses/sococ/mirage/
XV^{e} Ecole Européenne en E.D.A. Carcassonne, 8-13 septembre 2003.
http://www.unige.ch/ses/sococ/mirage/eeda.html

Pour un inventaire plus détaillé sur les ressources Internet voir :
BRINGÉ A. et LE GUEN M., (2002), « Ressources Statistiques via Internet, Un Aperçu », *Document de travail INED-MATISSE*, n°2002, 28 pages.
http://matisse.univ-paris1.fr/doc2/leguen1489.pdf

Pour accéder à des articles, prépublications et supports de cours sur l'Analyse Exploratoire par les Cartes de Kohonen, voir les pages de M. COTTRELL et P. LETREMY sur le site du SAMOS : http://Samos.univ-paris1.fr/